DOCUMENT RESUME

ED 147 343 TM 006 687

AUTHOR Myers, Charles T.
TITLE Test Fairness: A Comment on Fairness in Statistical Analysis.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RB-75-12
PUB DATE Apr 75
NOTE 12p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Analysis of Covariance; Equated Scores; *Mathematical Models; Statistical Analysis; *Statistical Bias; *Test Bias; Testing Problems; *Test Interpretation
IDENTIFIERS Equipercentile Equating Method

ABSTRACT

Fairness or unfairness may be an attribute of a test per se, or of its use, or of its statistical treatment. An hypothetical situation designed to be intrinsically fair and unbiased is used to show that analysis of covariance as a statistical method may introduce bias to the treatment of test scores. In contrast, equipercentile equating methods are shown, in this situation, to result in a fair and unbiased treatment of test scores. A graphic figure illustrates the comparison of the two different methods of analysis. (Author)

RESEARCH BULLETIN

RB-75-12

TEST FAIRNESS: A COMMENT ON FAIRNESS IN STATISTICAL ANALYSIS

Charles T. Myers

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
April 1975

Test Fairness: A Comment on Fairness in Statistical Analysis

Abstract

An argument is presented to suggest that the analysis of covariance may in some circumstances be an unfair method to use in the study of the question of test fairness. As an alternative, the use of equipercentile methods or equivalent linear methods may be preferred in these circumstances.

Test Fairness: A Comment on Fairness in Statistical Analysis

Fairness, like beauty, may well be in the eye of the beholder. There is no doubt that test fairness is difficult to define, to evaluate, or to prove or disprove. It may be a mistake to try to categorize a test or a test usage as either fair or biased. Instead a test or test usage should be evaluated as being either more or less fair than other available alternatives. Fairness in decision making, in an absolute sense, may be an impossible ideal. But in spite of all these difficulties and ambiguities, the maker and the user of tests is obligated to maintain the highest possible standard of fairness. There is also an obligation to clarify the meaning of the concepts of test fairness.

There have been a number of different and even incompatible definitions by such persons as Thorndike (1971), Darlington (1971), and Cole (Note 1) of what is meant by fairness, or conversely what is meant by bias in test scores. A distinction has been made by Flaugher (Note 2) between a biased test and the biased use of a fair test. This paper is an attempt to present a rationale for a fair analysis for determining whether a test is biased. What we intend to do is to describe data from a situation that appears to be intrinsically fair; and then we will compare two different statistical techniques for analyses of those data. We expect to show that the traditional technique may in some cases be intrinsically unfair and that the other technique may sometimes be preferable.

In a now famous study of test bias, Cleary (1968) said: "A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup." In this study she used the traditional regression method of analysis of covariance.

However, it may be instructive to consider a situation that can be assumed to be fair and then consider what would happen if we applied the analysis of covariance to data from that situation. Let us imagine a verbal aptitude test designed for use in fifth, sixth, and seventh grades and a parallel form of that test. Let us call these two tests Text X and Test Y. Let us assume that these tests are similar in content and, in the quality and difficulty of the items of which they are made up and that the test is equally appropriate for use in all three grades. Further, let us assume that Test X and Test Y have been carefully constructed so that any numerical score on Test Y is equivalent in meaning to the same numerical score on Test X. Under these circumstances, it seems reasonable to suppose that Text X is a fair test for predicting scores on Test Y.

Let us for convenience imagine that Tests X and Y have scores that range from 0 to 100 and that for either test the mean score for grade-seven children was 70, the mean score for grade-six children was 50, and the mean score for grade-five children was 30. Further imagine that the standard deviation of each within-grade distribution was 15, for each grade and for each test. Finally let us assume that the within-grade correlation between the two tests for each grade was 0.80. Obviously we are imagining hypothetical data simplified for the purpose of presenting a theoretical

position, but these hypothetical values are unrealistic because they are so regular, not because they are outside the normal range of common experience.

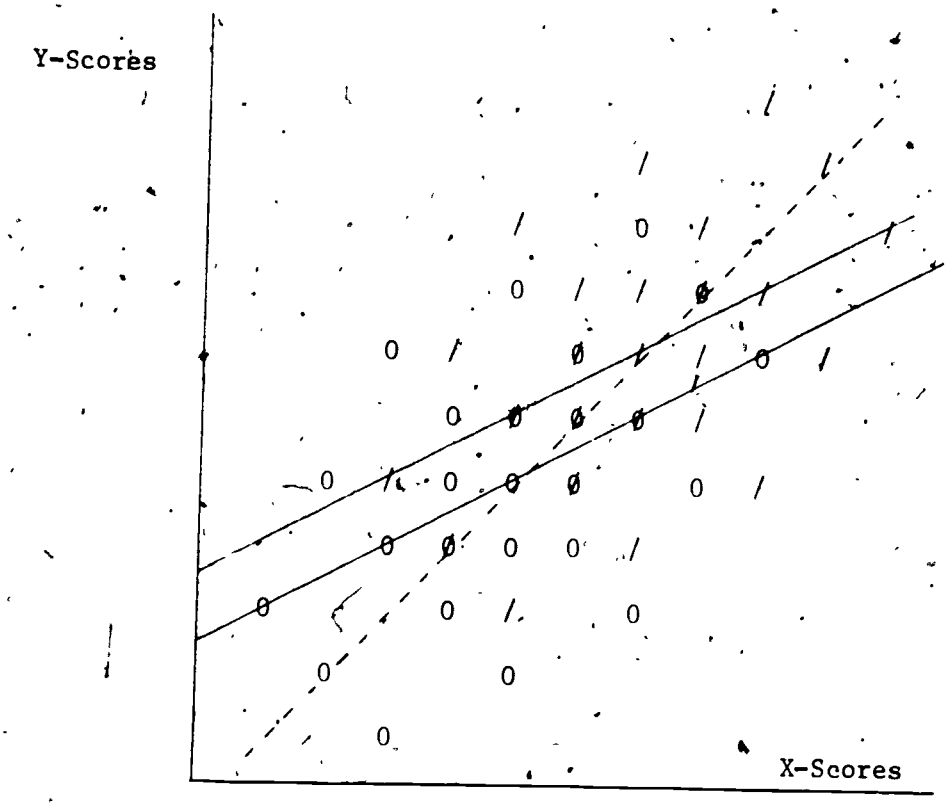
Given these conditions, consider what would happen if we applied analysis of covariance to the question: In comparison with its use for seventh graders, is Test X a fair test for fifth graders for the purpose of predicting scores on Test Y? Notice that we have supplied information to suggest that Test X and Test Y are identical in all the comparisons we have made and it seems that we may say intuitively that Test X is fair for that use. However, according to analysis of covariance, grade seven and grade five would not have identical regression lines. The regression lines would be parallel, but their Y intercepts would be different. Grade seven would have a Y intercept of 14 while grade five would have a Y intercept of 6, giving a difference of 8 points on the Y scale. The same difference would be found at other score levels. A fifth-grader with a score of 50 would have a predicted Y score of 46, but a seventh-grader with an X score of 50 would have a predicted Y score of 54. According to analysis of covariance as used by Cleary and several others, Test X may be considered biased against fifth-graders.

If a situation which was designed by definition to be fair is shown by analysis of covariance to be unfair, this suggests that perhaps analysis of covariance is inappropriate as a technique for studying this question. To make this point clearer, consider what would happen in this situation if we used Test Y scores to predict Test X scores for fifth-graders. Then we would find that Test Y was biased against fifth-graders in exactly the same amount. We have now reached the anomalous conclusion that both tests

are unfair and that both are biased in exactly the same way and amount in relation to the other.

As an alternative to analysis of covariance for studying test fairness, let us consider what would happen if we used equating or calibrating methods. Again let us disregard sampling errors and departures from linearity in order to clarify the analysis. For the kind of situation that we have described, the type of equating or calibrating most likely to be used would be equipercentile equating or the linear equivalent of setting means and standard deviations equal. From what we have been told about these two tests and the three grade groups, we would normally predict that (except for sampling errors) all three grades would show Test X as being equivalent to Test Y and therefore unbiased. The implication we draw from this analysis is that calibrating procedures are sometimes to be preferred to analysis of covariance in studies of test fairness or of test bias.

It may be advantageous to present these concepts graphically. In the figure below, two overlapping bivariate distributions are shown, with the members of the higher-scoring group indicated by /'s and the members of the lower-scoring group indicated by 0's .



In the figure above, the two slanting solid lines are the two regression lines and the dashed line is the equipercentile equating line. The lower of the two regression lines represents the regression equation for the lower-scoring group. In this illustration (which is admittedly hypothetical, but may be realistic) any particular X-score would be used to predict a lower Y-score for a member of the lower-scoring group than it would for a member of the higher-scoring group. In this particular illustration, the standard deviations for both groups on both variables are all equal, the means on both variables for the lower-scoring group are one standard deviation lower than for the other group, and the two within-group correlations are equal to 0.50.

This discussion has not presented a new concept of test fairness. The equipercentile relationship, or an equivalent linear relationship, has been discussed by Lord (1967), Thorndike (1971), Darlington (1971), and Myers (Note 3). What may be new in this context is the concept of proposing an inherently fair situation and considering what type of analysis would be logical to use for its evaluation; that is, the suggestion is to evaluate the statistical method by determining whether it might be expected to give a fair and unbiased answer.

There are two implications of this approach to the question of test fairness in comparison with the more traditional analysis of covariance approach. First, the use of this method would tend to be less likely to result in a decision that a test was biased against a lower-scoring group than would the analysis of covariance method. Second, the use of this method in admissions decisions would tend to result in more favorable decisions for the higher-scoring members of lower-scoring groups.

Although this illustration used the prediction of one test score by another, the model and principles may apply directly to the situation in which the Test Y of the illustration is replaced by some criterion performance such as grade-point-average in college or productivity on a job. But it is important to emphasize that this model is not appropriate to all such circumstances. For example, it would not be appropriate if the criterion itself were biased or irrelevant to the purpose of the test. In our illustration the two variables were equal in a number of ways, such as presumably equal in reliability, that would not commonly occur in a practical situation. The equipercentile model is no more of a panacea than is the

analysis of covariance. It is always important that the assumptions in the mathematical model should not be in violation of the facts of the particular situation, and whatever model is chosen must be appropriate to those facts.

Reference Notes

1. Cole, N. S. Bias in selection (ACT Research Report No. 51). Iowa City, Iowa: American College Testing Program, 1972.

2. Flaugher, R. L. The new definitions of test fairness in selection: Developments and implications (ETS RM 73-17). Princeton, N.J.: Educational Testing Service, 1973.

3. Myers, C. T. Bias and interpretation: Cases for ordinal measurement (ETS RM 73-18). Princeton, N.J.: Educational Testing Service, 1973.

References

Cleary, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Darlington, R. D. Another look at "culture fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Lord, F. M. A paradox in the interpretation of group comparisons. Psychological Bulletin, 1967, 68, 304-305.

Thorndike, R. L. Concepts of culture fairness. Journal of Educational Measurement, 1971, 8, 63-70.